Data Mining Ocean Model Output at the Naval Oceanographic Office Major Shared Resource Center

Pete Gruzinskas
Naval Oceanographic Office
1002 Balch Blvd.
Stennis Space Center, MS 39522
gruz@navo.hpc.mil

Andy Haas
Northrop Grumman Information Technology
1001 Balch Blvd.
Stennis Space Center, MS 39529
haas@navo.hpc.mil

Ludwig Goon
Northrop Grumman Information Technology
1001 Balch Blvd.
Stennis Space Center, MS 39529
lag@navo.hpc.mil

***Abstract*-One of the Computational Technology Areas supported by the High Performance Computing Modernization Program is Climate, Weather, and Ocean (CWO) modeling. To this end, state-of-the-art computing architectures are leveraged against the extremely difficult problem of mathematically modeling and predicting the behavior of a variety of ocean climatological parameters. The problem at hand is the technology to store, retrieve, manipulate, and display these data has not kept pace with the computational technology. During the last five years, we have seen significant cost reductions associated with applying the status quo in visualization techniques to scientific data sets. This is due in large part to the computer gaming industry, driven by the huge profit margins associated with that market. The scientific community has benefited by these advances in low-cost architectures, but only as a by-product of its original intent, which is entertainment. Even so, these low-cost architectures are not designed to handle the scale of data sizes presented by the scientific community and serve only to make inadequate techniques cheaper to field and use. The Naval Oceanographic Office Major Shared Resource Center (NAVO MSRC) Visualization Center is challenged with providing its users state-of-the-art analysis environments for the interrogation of their increasingly large data sets. This paper deals with the data generated by the CWO community, all of whom work with large domains and high resolutions (either vertically, horizontally, or both) that all vary over time. This leads to very large data sets (rows x columns x layers x attribute per cell) for each time step and can challenge even the most powerful architectures when trying to extract or "mine" information from the raw data. As in most visualization applications, the model output deals with physical parameters that are invisible to the naked eye. This means effective methods of display are required for ocean circulation or currents, sea surface height, temperature, salinity, and so on. One analogy, which no doubt started the concept of "data mining," is that the raw data represent a huge block of ore from which gold nuggets of valuable information (features) must be extracted or mined. This paper concerns the technical solutions that were built to solve the challenges described above, including algorithms, data descriptions, and formats.**

I. INTRODUCTION

The High Performance Computing Modernization Program (HPCMP) supports the Department of Defense (DOD) ocean modeling community by providing state-of-the-art resources to researchers involved in the extremely challenging task of mathematically predicting environmental events. These resources include time (cycles) at High Performance Computing (HPC) facilities, code optimization and parallelization, scientific visualization and analysis support, networking, and training. The program accomplishes this with a mix of four Major Shared Resource Centers (MSRCs) and 17 Distributed Centers (DCs), connected to each other via a high bandwidth (OC-48) Defense Research and Engineering Network (DREN). Information regarding our program or specifics about our center can be obtained at http://www.navo.hpc.mil. The HPCMP supports ten Computational Technology Areas (CTAs), but this paper will only deal with the Climate, Weather, and Ocean (CWO) modeling CTA. The science of ocean modeling on high performance computers has evolved into a multidisciplinary field. It combines the physical sciences associated with ocean modeling, along with the computer science needed to exploit the massively parallel computing systems like those located at the Shared Resource Centers (SRCs). All of this leads to massive model output, which has become increasingly difficult to manage and analyze. It appears the computational technology required to generate these massive fields of environmental data has outpaced the technology to store, retrieve, manipulate, display, and analyze these critical data. The role of the MSRC Visualization Center staff is to facilitate analysis of the critical data being generated daily on these HPC systems. A critical part of the Naval Oceanographic Office Major Shared Resource Center (NAVO MSRC) is its state-of-the-art archival facility, which can store petabytes of data in a fashion that allows users to access this information on demand.

One goal of the Visualization Center staff is to create a knowledge base from this archive of binary information. This involves far more than just visualizing the data. It must take into account optimal archive formats and associated metadata, along with software applications that not only create graphical structures from these data, but also provide some level of analytical function. This data management strategy will ultimately require some type of middleware that will facilitate intelligent retrieval via a browser or other common interface. The technology to accomplish this goal has come of age; it is a matter of developing and fielding these solutions in a manner that has minimal impact on our user base. One of the challenges associated with supporting our user base is the fact that they are geographically distributed. This places the computing resources and subsequent generated data in one place and the scientist who needs to analyze these data in another place.

One strategy to provide support to these geographically distributed users is the development of portable analysis applications. Although there is a plethora of commercial software, shareware, and freeware available that has a lot of data analysis functionality, there are problems with fielding solutions based on these software. In the case of commercial off-the-shelf (COTS) products, there are licensing costs, learning curves, and, frankly with the exception of one or two expensive applications, they are general purpose and can't deal with the data effectively. This prompted the NAVO MSRC Visualization Center staff to develop some custom OpenGL applications that create *interactive* analysis environments that provide a level of control over the modelers' domain, both spatially and temporally. These applications are portable and so effective they can be used to analyze global high-resolution ocean circulation model output on today's notebook computers. This does not put the analyst and data in the same place, but does facilitate analysis of data, which can be transferred (FTP) to the users location.

Another technology called "remote rendering" is being fielded to help alleviate the painful process of transferring huge time-series data. Some of these ocean circulation models generate time steps that approach 1 Gigabyte in size! To effectively analyze these data, long time scales on the order of several months to a year need to be assembled to feed these custom applications. These applications basically regenerate the modelers' computational domain in a virtual environment. Remote rendering allows geographically dispersed users to leverage high performance visualization servers located at the MSRCs. These servers not only have the horsepower to manipulate the large ocean circulation model output, but also have large storage partitions connected with ultra-fast fiber channel interfaces to stage these data. Remote rendering leverages the high-speed networks put in place by the HPCMP along with ever-increasing bandwidth to the desktop. It performs the entire rendering process on the remote server and passes just the output of the frame buffer (pixels) across the network to the user. Various compression algorithms are available to the user as well as an interface to add custom compression schemes. The commercial product selected for this initiative is SGI's Vizserver application. At present SGI has client versions that will support IRIX (SGI), Solaris (Sun), and Linux (PC Unix), platforms. We are also evaluating Hummingbird's X3D, which employs a distributed rendering model.

II. DATA MINING OVERVIEW

Over the last three to four years, the NAVO MSRC Visualization Center has built a "virtual toolbox" of techniques and algorithms geared toward the analysis of 4-D data grids. More significantly, we have postured the center to deal with extremely large data sets. We've accomplished this with hardware, software, and networking. Specifically we use SGI's Onyx2 graphics server with close to a terabyte of fiber channel raid. This server has 8 processors, 8 gigabytes of RAM, and 64 megabytes of texture memory. It also has ATM OC-12, as well as Gigabit Ethernet network interfaces. Recent tests using GigE with "jumbo frames" between our graphics server and our mass storage archive server showed we could transfer (KFTP) 1 gigabyte files in approx 25 seconds. We have installed gigabit Ethernet fabric throughout the Visualization Center to the DREN backbone. Our desktop environment has been upgraded to powerful dual-Pentium machines with a gigabyte of RAM and NVIDIA graphics. The center runs a combination of Windows and Linux, with Linux providing a far greater level of security and capability.

We have explored a variety of techniques for the analysis of ocean circulation and are working with others in academia and industry on new techniques and better ways to apply old techniques. I will briefly discuss a few of these approaches, but the following section will outline what we consider to be a major breakthrough toward bringing terrascale data to the desktop for analysis. We are currently in the patent process for some of our tools, which will not be covered in this paper.

An interesting approach was taken by rendering surface currents as a volume of time (Fig. 1). What this means is slices of scalar surface currents were accumulated and rendered as a volume. The Z component of the volume is time, so interesting temporal features, such as tidal constituents, become easily visible as corkscrew-type features extending through the volume.

Another interesting technique applied to vector current data is called the colorwheel (Fig. 2). We learned about this technique from Mississippi State University (MSU). MSU is one of the academic institutions that provides support to the HPCMP through its Programming, Environment, and Training (PET) component. The talents in this field among our academic institutions is no secret, and this program has a focused effort to leverage this expertise. The colorwheel method of 2-D vector field visualization uses the Hue-Saturation-Value (HSV) color space to map vector direction to Hue (color) and magnitude to Saturation and Value. This visualization technique is better than traditional primitive graphics (arrows) in high-resolution models because it offers less visual congestion. In addition, the colorwheel technique displays better eddy formation and vector magnitudes in circulation models.

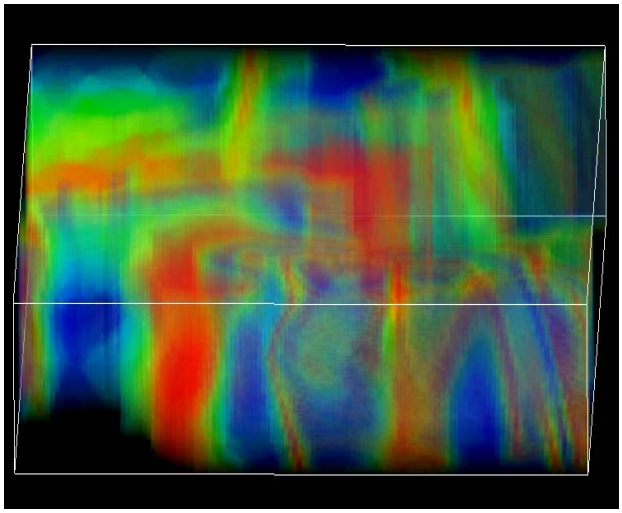

Fig. 1. Perspective view of time-series surface currents.

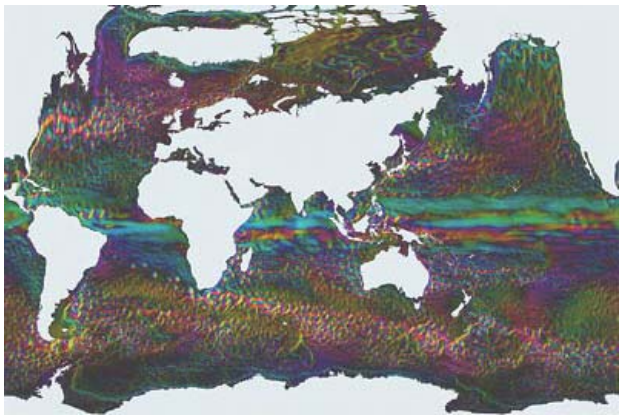

Fig. 2. Colorwheel display of vector data.

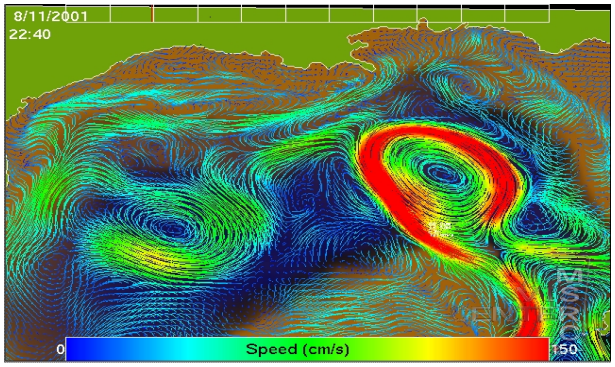

Fig. 3. Pathlines of PDOM surface currents.

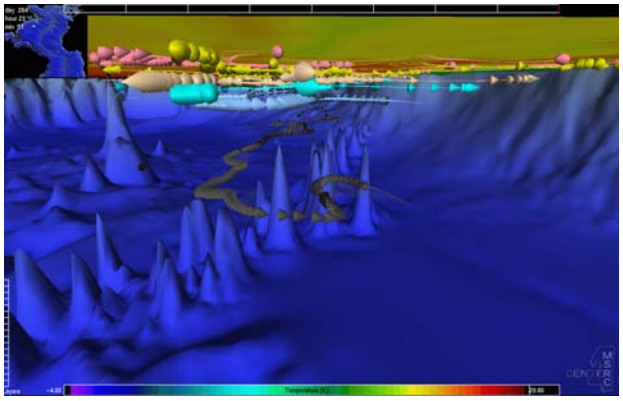

Fig. 4. Pathlines of MICOM's 3-D current field.

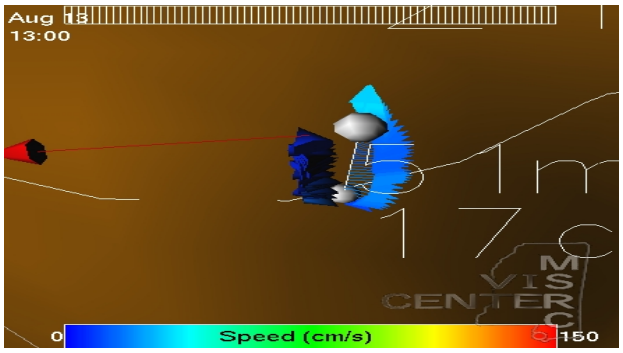

Fig. 5. ADCP buoy next to SWAFS probe.

As mentioned earlier, analytical software applications must do more than simply display the model output. Algorithms must be applied, which can provide some level of analysis to the modeler's domain. One such example is an application that applies advection algorithms to the time-varying current field. This technique is analogous to and offers a direct comparison to Lagrangian drifter data.

Our applications allow the user to "seed" particles anywhere within the model domain, horizontally and/or vertically, and observe their movement as they advect over time (Figs. 3 and 4). This has proven to be a powerful analysis tool. A significant application of this environment was the application built to analyze the current field in the vicinity of the Japanese fishing trawler *Ehime Maru*. The display of an Acoustic Doppler Current Profiler (ADCP) was incorporated into this virtual environment to provide a direct comparison between measured currents and model output (Fig. 5). A useful feature incorporated into this and most of our custom applications is the playback function. This allows the analyst to record a session for playback at a later time, with the same data or other data. It is essential for reproducing the results of a particular analysis session. The application merely writes a small binary file of mouse and keyboard events, which it can subsequently read and execute. This is also very useful for demonstrations as it eliminates the need for a "driver."

A recent effort to ground truth ocean model output involves the comparison of the model Sea Surface Temperature (SST) field with measured SST information gathered from National Oceanic and Atmospheric Administration's Multichannel Sea Surface Temperature (MCSST) satellite (Fig. 6). We routinely receive the MCSST information from the Naval Oceanographic Office (NAVOCEANO) Warfighting Support Center (WSC), and after the data are converted to a network-based common data format (netCDF), it is subtracted from the model SST. On-the-fly interpolation is applied to account for varying resolutions. The resolution of the MCSST data, which are collected by polar-orbiting satellites employing Advanced-Very-High-Resolution Radiometers (AVHRR), is 3600 X 1800.

III. RANGERSCOPE

The three basic model types we work with are Isopycnal, fixed-Z layer depth, and sigma. The fixed-Z models are the Parallel Ocean Program (POP) and Princeton Dynamic Ocean Model (PDOM). The XY grid for POP is irregular, while PDOM has a regularly spaced grid. The isopycnal models are the Navy Layered Ocean Model (NLOM) and the Miami Isopycnal Coordinate Ocean Model (MICOM). The XY grid for NLOM is regularly spaced. The spacing for MICOM is regular in longitude and irregular in latitude (with the spacing between latitudes decreasing as they run north). The sigma models are typically much smaller in size than that of their large isopyncal and fixed-Z counterparts. They are generally basin-scale models with smaller, often irregular grids that match a specific area of interest (e.g., Persian Gulf). We have focused data mining to the larger models as they have

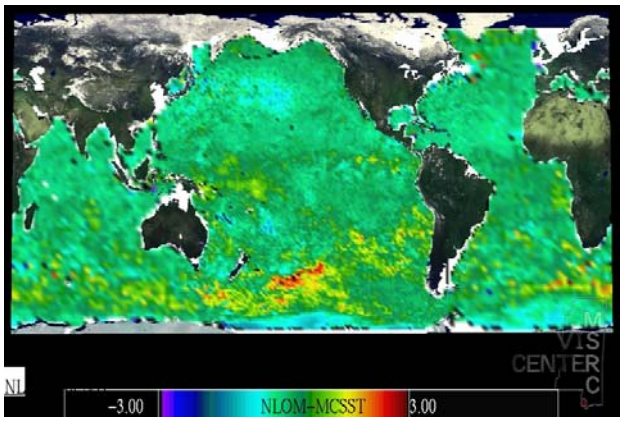

Fig. 6. RangerScope display of NLOM-MCSST residual.

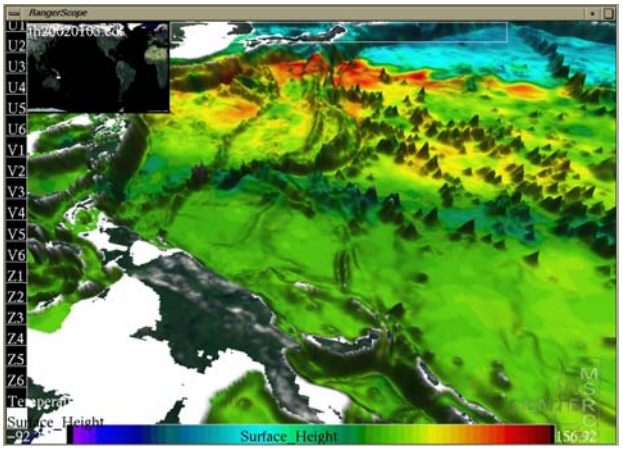

Fig. 7. Perspective view of NLOM Sea Surface Height draped over bathymetry.

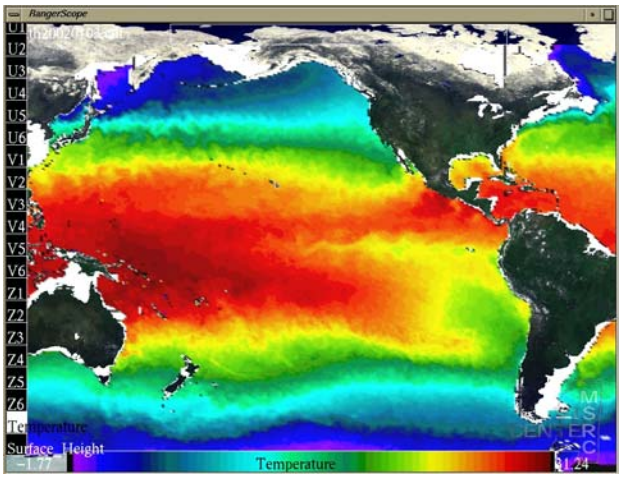

Fig. RangerScope display of NLOM sea surface temperature. Note: sidebar showing all model output fields available for display.

large arrays of several data attributes spread across many vertical layers. We also describe how a standard netCDF is used to store the data arrays.

There are two classes of data mining applications: one explores global data in its native format at interactive fashion, and the second applies interactively driven advection analysis to local areas of the model.

An application called RangerScope satisfies the first mining class, which we shall focus on in this paper. It is a general-purpose tool for mapping sequences of large 2-D data files onto an optional 3-D terrain elevation (Fig. 7). The tool enables the user to roam across the field of data while it is played back at interactive speeds. The data are kept in a sequence of netCDF files. Any 2-D variable data set present inside the files is eligible for viewing by clicking the mouse. Any number of 2-D variables inside any number of files is accommodated. The resolution of each 2-D variable may be very high (the largest case has been a 8000 x 4000 gulf-coast bathymetry). The tool uses a dual parent/child process and shared memory implementation on a Unix/Linux system. The child process performs the data input/output to support the animations. The parent renders the current set of data while the child prepares the next set.

RangerScope is able to render very large data arrays because it uses dynamic level-of-detail algorithms. These algorithms enable the user to travel real-time anywhere in the field and see information at its native resolution. The algorithms have common applications across all cases where large data arrays need to be managed for interactive exploration.

Any 2-D netCDF variable can be displayed by itself (flat) or over a 3-D terrain. The optional terrain data is given as a separate variable in a netCDF file. The variable's values are treated as Z (elevation) data. These Z values can have different resolutions and XY grid locations than the animated 2-D data variables. The 2-D data grid is correlated with the Z-grid so that the data's location matches with the terrain's. A land mask is created that tells which areas of the Z-grid are above sea level. The land mask can be drawn either as a solid color or with a given RGB image file as a texture. A texture image allows land to be colored with high-resolution satellite imagery of terrain features such as mountains, plains, forests, and rivers. Users have found this feature useful in correlating ocean patterns near land with topological features such as river outlets.

RangerScope has broad applications for all types of terrain and data. It has been used to analyze 1/10 degree global bathymetry mapped with MCSST. Variables containing the difference between model (predicted) temperatures versus MCSST (actual) temperatures have been produced into a time sequence of netCDF files. The POP, NLOM, and MICOM are key models where research between model versus real are taking place.

The POP model grid has 40 layers each at a fixed Z-depth. The X and Y grid resolution is 3600 x 2400. The XY grid becomes highly irregular as it approaches the North Pole. The irregularity was designed to eliminate the polar singularity of a regular longitude-latitude configuration. Because the layers are fixed-Z, the number of layers containing valid data differs across the bathymetry. Each grid node has temperature, salinity, and 3 directional velocity components. Layers of components (in this case the top 20 layers) are saved as separate 2-D variables in netCDF files. Any variable can be navigated and displayed across a time sequence of netCDF files via the RangerScope application.

The NLOM is a 1/16 degree (4096 x 2304) regularly spaced longitude, latitude grid (Fig. 8). It is a deep-water model evaluated at areas of the ocean at least 200 meters deep. It uses six isopycnal layers to represent the stratification of water density. The isopycnal nature varies the thickness of each layer across the grid. Because the thickness changes over time as well, each grid node has a Z depth component. NLOM outputs U and V velocity and Z depth values for each layer. The surface temperature and sea height are also output for a total of 20 2-D variables (6 U's, 6 V's, 6 depths, plus 1 temperature and 1 height). All 20 variables are labeled and packaged inside a netCDF file. NLOM outputs one netCDF file per 24-hour time step.

MICOM covers only the North Atlantic. The longitudes are spaced uniformly, while the latitudes run rectilinearly so the values are spaced closer farther north. MICOM has 16 isopycnal layers containing U, V, and Z values at each node point.

Applications for particle advection have been developed for all of the ocean circulation models described under that section. The applications use the U, V, and Z data in the model's netCDF files to reconstruct the flow patterns within selected layers. Particles in the flow are advected on the fly over a continuous time interpolated series of daily output files. The user is able to draw the placement of particles in any area of the model across the full scale of vertical layers.

V. FUTURE WORK

We are currently porting the applications described herein to the Hybrid Coordinate Ocean Model (HYCOM). As its name suggests, it is a hybrid configuration that uses isopycnal layers in the open ocean, sigma layers in the near shore areas, and fixed-Z levels in the mixed layer.

Remote rendering will play a critical role in the dissemination and analysis of high-resolution model output. The NAVO MSRC Visualization Center staff will continue to evaluate technologies that reduce the limitations on analysis created by physical distance. This includes collaborative or data-sharing technologies that allow disparate groups or individuals to view and analyze the same domain simultaneously.

The HPCMP supports atmospheric modelers as well. A coupled display in lieu of a true coupled model may serve as a valuable diagnostic tool for both atmospheric and oceanic circulation models.

We need to document these accomplishments and applications more adequately and make this information available on our web site. We believe that these tools are unique and of much value to the ocean modeling community. They can be extended beyond ocean model analysis to whatever data the 4-D grid represents. So many pixels, so little time.

Acknowledgments

Since we discussed our work with the Shallow Water Analysis and Forecast System (SWAFS) at Oceans '99, we have expanded our support to include many circulation models, from both the DOD and the Department of Energy. The following is intended to identify some of the researchers with whom we have collaborated and hopefully serves to promote the excellent work they have accomplished in this field. Without their efforts our work would not have been possible or necessary.

SWAFS is a basin-scale sigma-level Navy operational circulation model. Our initial development work was conducted with Dr. Martha Head of NAVOCEANO. Follow-up work, particularly for studies surrounding the retrieval of the Japanese fishing trawler *Ehime Maru*, was done with Dr. Charles Horton and Melody Clifford of NAVOCEANO.

NLOM is a global 1/16 degree, 6-layer isopycnal ocean circulation model. It has recently been transitioned to operational status for the U.S. Navy. Development efforts were conducted with Drs. Alan Wallcraft, Jay Shriver, and Ole Martin Smedstad.

MICOM is an isopycnal ocean circulation model, which covers the North Atlantic Ocean. Drs. Eric Chassignet and Zulema Garraffo at the Rosenstiel School of Marine and Atmospheric Science University of Miami are our collaborators on this effort.

POP is a global 1/10 degree 40-layer fixed-z layer global ocean circulation model that is designed to represent the ocean component of a coupled ocean, atmosphere, and ice model. We have worked with Dr. Julie McClean and Detelina Ivanova at the Naval Postgraduate School (NPS) on this application development effort.

The Polar Ice Prediction System is a POP-based ice drift model. We worked with Dr. Wieslaw Maslowski at the Naval Postgraduate School on this project.

The Multi-block Grid Princeton Ocean Model (MGPOM) is a 3-D (U, V, W) ocean circulation model built around a multi-block grid approach. We collaborated with Dr. Phu Luong at the Engineering Research and Development Center on this development effort.

PDOM is a high-resolution, 37 fixed-Z layer, basin-wide circulation model for the Gulf of Mexico. We worked with Dr. John Blaha and Pat Wilz of NAVOCEANO on application development for this circulation model.

The Massachusetts Institute of Technology Ocean General Circulation Model (MITgcm) is a coarse-resolution (1 degree to 1/3 degree), 46-layer ocean circulation model. These data mining efforts were conducted with the assistance of Zhangfan Xing at National Aeronautics and Space Administration Jet Propulsion Laboratories.